p.1

Posterior distribution

$$p(w \mid X, t)$$ $$t = y + \varepsilon = Xw + \varepsilon$$

$$X = \begin{bmatrix} x_1^{(1)}, & \dots, & x_1^{(D)} \\ x_2^{(1)}, & \dots, & x_2^{(D)} \\ & & \\ x_N^{(1)}, & \dots, & x_N^{(D)} \end{bmatrix} \Rightarrow \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \qquad X : N \times D$$

$x_1 = [x_1^{(1)}, \dots, x^{(D)}]^T$: D-dimensional input

$w = [w^{(1)}, \dots w^{(D)}]^T$: D-dimensional parameter vector

$t = [t_1, \dots, t_N]^T$: N data samples (here: output), each output t_n is one-dimensional, N samples of one-dim output forms a N-dimensional column vector

We expect our posterior to follow a Gaussian distribution:

$$p(w \mid X, t) = \mathcal{N}(\mu_w, \Sigma_w)$$

μ_w - mean
Σ_w - covariance of the W posterior

$$\frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad p(w \mid X, t) \sim e^{-\frac{1}{2}(w-\mu_w)^T \Sigma_w^{-1} (w-\mu_w)} =$$

$$= \underbrace{e^{-\frac{1}{2} w^T \Sigma_w^{-1} w}}_{\text{quadratic}} \cdot \underbrace{e^{w^T \Sigma_w^{-1} \mu_w}}_{\text{linear term}} \cdot \underbrace{e^{-\frac{1}{2} \mu_w^T \Sigma_w^{-1} \mu_w}}_{\text{const.}}$$

We are going to "complete the square" soon to identify μ_w and Σ_w.

Likelihood

$$\underline{p(t|w, X)}$$

$$t = y + \varepsilon = Xw + \varepsilon$$

$$p(t|w, X) = \mathcal{N}(Xw, \sigma^2 I_N)$$

N-dimensional Gaussian distr.
with a diagonal covariance
(conditional independence of t_n)

Prior over w

$$p(w) = \mathcal{N}(0, \Sigma^{-1})$$

So, posterior $\quad p(w|t, X) \propto p(t|w, X) \cdot p(w)$

$$\Downarrow$$

$$p(w|t, X) \propto e^{-\frac{1}{2\sigma^2}(t - Xw)^T (t - Xw)} \cdot e^{-\frac{1}{2} w^T \Sigma^{-1} w}$$

If we focus now on the exponent:

$$-\frac{1}{2\sigma^2}(t - Xw)^T (t - Xw) - \frac{1}{2} w^T \Sigma^{-1} w =$$

$$= \underbrace{-\frac{1}{2\sigma^2} t^T t}_{\text{const. (from } w \text{ perspective)}} + \frac{1}{\sigma^2} t^T X w - \frac{1}{2\sigma^2} (Xw)^T (Xw) - \frac{1}{2} w^T \Sigma^{-1} w$$

p.3

We have then 3 terms:

(I) $-\frac{1}{2\sigma^2} t^T t \rightarrow$ const. for w

(II) $\frac{1}{\sigma^2} t^T X w = \frac{1}{\sigma^2} w^T X^T t \rightarrow$ term linear in w

(III) $-\frac{1}{2\sigma^2} (Xw)^T (Xw) - \frac{1}{2} w^T \Sigma^{-1} w =$

$$= -\frac{1}{2\sigma^2} w^T X^T X w - \frac{1}{2} w^T \Sigma^{-1} w =$$

$$= -\frac{1}{2} w^T \left(\frac{1}{\sigma^2} X^T X\right) w - \frac{1}{2} w^T \Sigma^{-1} w =$$

$$= \underbrace{-\frac{1}{2} w^T \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1}\right) w}_{\text{quadratic term in } w}$$

If we look back at our posterior $p(w|t,X)$ we can identify corresponding terms.

Let's complete the square:

(I) let's leave out corresponding const. terms $-\frac{1}{2\sigma^2} t^T t$

(III) quadratic terms:

$$-\frac{1}{2} w^T \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1}\right) w$$

vs.

$$\cancel{-\frac{1}{2} w^T \Sigma^{-1} w} \quad -\frac{1}{2} w^T \Sigma_w^{-1} w$$

So, $\Sigma_w^{-1} = \frac{1}{\sigma^2} X^T X + \Sigma^{-1}$

II linear terms

$$\frac{1}{\sigma^2} w^T X^T t \quad \text{vs.} \quad w^T \Sigma_w^{-1} \mu_w$$

$$\Downarrow$$

$$w^T \Sigma_w^{-1} \mu_w = w^T \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1} \right) \mu_w$$

$$\Downarrow$$

$$\cancel{w^T} \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1} \right) \mu_w = \frac{1}{\sigma^2} \cancel{w^T} X^T t$$

$$\underline{\underline{\mu_w = \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1} \right)^{-1} X^T t}}$$

So, let's summarise the posterior

$$p(w \mid t, X) = \mathcal{N}\left(\frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1} \right)^{-1} X^T t, \; \frac{1}{\sigma^2} X^T X + \Sigma^{-1} \right)$$

σ comes from the likelihood

Σ comes from the prior (covariance)